2014

IMPACT OF NON-INFORMATIVE ALTERATION IN ASSESSMENT: IDENTIFICATION AND CORRECTIVE PROCEDURES

Aaron M. Baker
University of Rhode Island, aaron.m.baker@gmail.com

IMPACT OF NON-INFORMATIVE ALTERATION IN ASSESSMENT: IDENTIFICATION AND CORRECTIVE PROCEDURES

BY

AARON M. BAKER

A THESIS SUBMITTED IN PARTIAL FULLFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

PSYCHOLOGY

UNIVERSITY OF RHODE ISLAND

2014

MASTER OF ARTS THESIS

OF

AARON M. BAKER

APPROVED:

Thesis Committee:

Major Professor David Faust

Grant Willis

Lisa Weyandt

Leslie Mahler

Nasser H. Zawia
DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2014

ABSTRACT

The accuracy of psychological and neuropsychological evaluation may be degraded when non-informative alteration (NIA) in data, or alteration in the manner data are represented but not in core meaning, impacts interpretation. If NIA exerts an impact, it may also lead to underutilization of truly useful information. Certain interpretive practices that are based on configural relationships and are already problematic (e.g., scatter analysis) may be particularly vulnerable to NIA and thereby further compromised. This study examined: (1) judgments regarding inter-test scatter across a neuropsychological battery, (2) whether NIAs impact judgments regarding scatter, and (3) whether truncating the visual presentation of scatter alters misjudgments about the frequency or pathological significance of scatter. Participants (N = 193) were neuropsychologists and graduate students who have received training in neuropsychological assessment. When judging neuropsychological profiles, participants markedly overperceived normal levels of scatter as rare or aberrant. The influence of NIA was mixed. Changing the visual plotting of percentiles from equal- to unequal-sized units did not alter judgments. In contrast, simply changing the designated metric from percentiles to T-scores, while holding visual plotting constant, reduced overperception of scatter, although only partially or insufficiently. An intervention that truncated visual scatter further improved judgmental accuracy (i.e., truncated visual scatter compared to larger visual scatter with mathematically identical information attenuated misjudgments about the normality of scatter). This study provides preliminary evidence for a previously underidentified source of error in the interpretation of psychological test data. Future research should determine whether the

current findings can be replicated, advance the design of interventions as needed, and assist in developing evidence-based standards for representing graphical displays that diminish the influence of NIAs.

ACKNOWLEDGMENTS

I want to recognize and send my sincerest gratitude to my mentor and major professor, David Faust, for his endless patience and support. In particular, I thank him for his unyielding moral convictions and commitment to intellectual pursuits. His mentorship is invaluable to my professional development. I also thank the other members of my thesis committee: Grant Willis, Lisa Weyandt, and Leslie Mahler. This project is successful thanks only to the breadth of expertise bestowed generously by my committee. My committee chair, Arthur Mead, deserves special thanks. His insight and support are greatly appreciated.

I am most grateful to my wife and family for their ongoing and unconditional support. My wife, Sarah, has provided me the encouragement and modeling of professional practice for which I am forever indebted. Lastly, my parents and three siblings offer a genuineness that is unmatched. The collection of each of their individual accomplishments will always bolster my own pursuits.

TABLE OF CONTENTS

LIST OF TABLES

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

CHAPTER 1

INTRODUCTION

The accuracy of psychological and neuropsychological evaluation may be degraded when non-informative alteration (NIA)[1], or alteration in the manner data are represented but not in core meaning, impacts interpretation. Additionally, psychologists might underutilize diagnostic, or truly useful, information, when NIA exerts an impact. It is posited that NIA in the presentation of psychological test data, such as changes in the physical dimensions of displays or even in the labeling of metrics, may nevertheless influence interpretation, therefore increasing the potential for error. Differences in visual presentation of test data occur with great regularity. As further explained below, examples of such differences include: denotation metric, graphical dimensions, coloration, and orientation. A scientific basis to evaluate and, where needed, reduce or eliminate impact from NIA is lacking.

Certain interpretive practices that are based on configural relationships (e.g., scatter analysis) may be particularly vulnerable to influence from NIA. When such alterations degrade the accuracy of certain interpretive practices, corrective procedures for graphically displaying data could prove highly beneficial. The present study explored: (1) judgments regarding inter-test scatter across a neuropsychological

[1] For the purposes of this thesis, "non-informative alteration" refers to any alteration in information that has no true value. Although NIA can come in different forms, of particular interest here are changes in the graphical display of data that in no way alter the mathematical properties of those data. For example, if a scaled score of 80 represented in green color ink is changed to blue color ink, the change would be a NIA as defined here.

battery, (2) whether NIAs impact judgments regarding scatter, and (3) whether truncating the visual presentation of scatter attenuates or corrects misjudgments about the normality of scatter.

CHAPTER 2

REVIEW OF LITERATURE

Maximizing accuracy when applying psychometric measures depends on adherence to certain principles. Following sound procedures in selecting, administering, and interpreting tests are paramount in achieving accuracy (Mitrushina, Boone, Razani, & D'Elia, 2005; Strauss, Sherman, & Spreen, 2006). In regards to interpretation, properly developed and implemented statistical judgment methods are not influenced by NIA, and therefore, the accuracy of such methods should not be impacted. However, these methods are apparently underutilized (Vrieze & Grove, 2009). By contrast, subjective or clinical judgment may be susceptible to NIA. Clinical judgment may be influenced or degraded by factors that are non-informative secondary to cognitive limitations and biases (Faust, 1984, Faust & Ahern, 2012; Wedding & Faust, 1989). For example, the salience of information, even if unrelated to its diagnostic value, may heavily influence impressions and decisions. When interpretation is compromised by NIA, which may in turn lead to underutilization of diagnostic information, the accuracy of clinical decision-making is likely to decline.

Clinical interpretive practices are highly variable. Whether preferable or not, clinical or impressionistic judgment remains the most frequently used method for drawing conclusions and predicting outcomes across many domains of applied psychology (Vrieze & Grove, 2009) and has been promulgated as a core feature of neuropsychological test interpretation (Lezak, Howieson, Bigler, & Tranel, 2012).

Given the frequent use of clinical judgment in psychological and neuropsychological assessment, the potential impact of NIA should be examined and, if and when present, attenuated or eliminated to the extent possible. The present study focuses on whether holding data mathematically constant, but varying the manner in which it is represented visually, alters interpretation. The variations examined are intended to exemplify the range of common practices in the field of psychological testing and among test publishers.

Non-Informative Alterations

Within the field of psychological assessment, raw test scores are most frequently converted to standard scores (e.g., z-scores, T-scores, and percentiles) to aid in comparison with normative groups and from test to test. Test developers and publishers vary in the selected metric, some emphasizing T-scores, some Wechsler-like standard scores (e.g., mean = 100; standard deviation = 15), and some percentiles. Utilizing the same underlying data, different metric selection will often result in graphical displays that differ in appearance. Assuming a normal distribution for all measures, metrics are easily transformed into one another (although relationships may not be linear, e.g., percentiles to scaled scores). Therefore, substantively discriminating information should usually remain unchanged or identical across metrics.

There has been considerable debate within the field concerning the most useful metrics for representing test data. For example, some researchers adamantly oppose the use of percentiles (Bowman, 2002), and others endorse their value (Crawford & Garthwaite, 2009; Crawford, Garthwaite, & Slick, 2009). One limitation of percentiles

is the lack of interval measurement, resulting in disparity in the relationships between test scores (i.e., the relationship between the 10th and 20th percentile is disproportionate to the relationship between the 40th and 50th percentile). Therefore, graphically representing percentiles as equal-sized units, a mathematical distortion, visually skews the relationship between data points. At times, test data presented as percentiles are represented in unequal-sized units, consistent with their mathematical properties, but at other times, they are represented in equal-sized units, despite the resultant distortion.

For example, for the *Woodcock–Johnson III Tests of Cognitive Abilities* (Woodcock, McGrew, & Mather, 2001) results are graphed horizontally as unequal-sized units (i.e., proportional to a normal distribution); whereas results for the *Wechsler Memory Scale –III (WMS-III)* Auditory Composite Index (Wechsler, 1997a) are graphed vertically as equal-sized units (i.e., equal spacing between percentile points). Using equal interval spacing introduces increased visual discrepancy when plotted in percentiles and can lead to marked alterations in visual displays. Such alterations also occur with other metrics. For example, T-scores can be displayed to adhere to a normal distribution (i.e., equal-sized units), as is the case for the *Minnesota Multiphasic Personality Inventory-2* (*MMPI-2*: Butcher, Graham, Ben-Porath, Tellegen, Dahlstrom, & Kaemmer, 2001), or as unequal-sized units (Bowman, 2002). T-scores graphed at unequal-sized units are often visually similar to percentiles graphed as equal-sized units.

Not only does metric selection and interval spacing impact graphical display, but dimensions and orientation of a graph also regularly vary across, and sometimes

even within, psychological tests. With computerized-based test interpretation (CBTI), marked variations may occur even for the same test. For example, multiple commercial CBTI programs are available for the MMPI-2. Graphical displays on these programs use different dimensions and interval spacing to present T-scores (Williams & Weed, 2004a; 2004b). For example, portrait versus landscape orientation of the graphical display of test data can change the physical space between higher and lower scores and consequently visual impressions about test score disparities. Thus, regardless of psychologists' positions on the advantages and disadvantages of competing metrics, the graphical displays that everyday practitioners use often vary independently from the properties of the obtained test data themselves.

Scatter Analysis as a Susceptible Interpretive Practice

The potential problem of inconsistent interpretation based solely on variation in visual representation of test data may be substantial and pervasive within psychological assessment. For example, clinical interpretive practices often rely on scatter – relative variability between high and low scores. Interpretation of scatter attends to the relationship between patterns of high and low test scores and comparison of such to expectations about *normal* vs. *abnormal* test performance. Of concern, the visual distance between test scores is partly determined, and may be subsequently altered, by interval spacing, metric selection, and graphical dimensions. Although limitations in scatter analysis have been recognized for over half a century (Schofield, 1952), the appraisal of intra- and inter-test scatter remains one of the most common approaches to the psychological evaluation of cognitive function and brain disorders (Lezak, Howieson, Bigler, & Tranel, 2012).

Clinicians frequently underestimate normal levels of scatter (Schretlen, Munro, Anthony, & Pearlson, 2003), leading to overidentification of pathology. Schretlen and colleagues' study (2003) demonstrated that the great majority of normal adults show *marked quantitative discrepancies* in test scores across a prototypical neuropsychological battery, levels that are often considered deviant. Variability between test scores is often influenced not only by true differences in level of ability but by measurement artifact as well, such as the number of tests administered (Binder, Iverson, & Brooks, 2009), scoring errors, which may occur with surprising frequency (Allard & Faust, 2000; Simons, Goddard, & Patton, 2002), and inadequate normative selection (Brooks, Strauss, Sherman, Iverson, & Slick, 2009). Therefore, graphical presentation that accentuates differences between scores may compound these interpretive problems. If NIA worsens an already common, problematic judgment practice, the impact may be pervasive.

The primary concern here is broader than judgments about test scatter, as many facets of test interpretation may be influenced by alteration in graphical displays (e.g., impressions about the extremity or abnormality of test results). However, because of the frequent use and potential relevance of scatter analysis, it seemed an appropriate starting point for the study of the more general issue of graphical display. This thesis aimed to examine whether NIA may worsen an already problematic judgment habit, and, if so and better understood, might inform attempts to develop alternative or corrective methods of displaying data.

Pilot Study

Given the lack of research in this area, a pilot study was performed. The pilot study explored the potential impact of altering graphical displays of identical test data on psychological test interpretation. The study used a between-group design with one independent variable (graphical display plotted as one of two metrics: Wechsler standard scores [SS] or percentiles with equal-sized units [PES]) and two dependent variables (perceived frequency and level of aberrance). Participants (N = 11) consisted of graduate students in clinical or school psychology who had just completed a cognitive assessment course and, therefore, had recent training in the interpretation of the *Wechsler Adult Intelligence Scale – IV* (*WAIS-IV*: Wechsler, 2008). This test generates a number of summary or composite scores. Participants were randomly assigned to one of two groups: SS (n = 5) and PES (n = 6). Each participant reviewed a graphical display of composite scores from one prototypically normal *WAIS-IV* profile. Participants were then asked to judge how frequently the level of scatter occurs in the normal population and the degree to which it might indicate aberrance. Judgments regarding frequency and aberrance may appear to be addressing the same question, but this is not necessarily the case. For example one might judge test results as rare, but not indicative of dysfunction (or vice-versa).

Although the pilot study used the *WAIS-IV* (Wechsler, 2008), the prototypically normal profile was developed using the normative database from the Wechsler Adult Intelligence Scale – III (*WAIS-III*: Wechsler, 1997b) based on Ahern, Faust, and Bridges' work (in preparation). Using the intercorrelations and base rate discrepancies between IQ and index scores for the normative sample set forth in the

WAIS-III – WMS-III Technical Manual (2002a), Ahern et al. created a profile that was normal or unremarkable in all basic respects (e.g., level of scatter among composite scores).

Revisions from the *WAIS-III* to the *WAIS-IV* required minor adjustments in the prototypically normal profile. More specifically, where applicable, the labels for new or revised composite scores were substituted for previous labels, and extrapolations were made about generalization of normal features across versions of the test given similarities in key psychometric properties (e.g., generally satisfactory levels of test reliability, common means and standard deviations). There are limitations to such extrapolation, e.g., correlations between indexes and IQ scores vary between the *WAIS-III* and *WAIS-IV*. However, for the purposes of the pilot study, which aimed merely to explore a potential influence of NIA, the impact of such limitations on the overall results was viewed as likely to be minor. Whether the profile was prototypically "normal" or "abnormal" the data were mathematically constant, and differences in judgments of frequency or aberrance would presumably be due to NIA.

The prototypical profile was plotted in one of two metrics: SS or PES (see Figure 1). The PES profile introduced greater visual scatter compared to the SS profile, although both are mathematically constant with one another. No additional information was provided beyond the visually represented test scores. Participants were instructed as follows:

> Please respond to the questions below concerning the frequency and overall aberrance, if any, in the discrepancy between Wechsler IQ scores and Indexes. It is understood that in clinical practice the information provided is insufficient

and other relevant information would be advantageous, e.g., referral question, demographic information and clinical history. Given the acknowledged limits of the information, please respond as best you can.

As previously noted, the dependent variables included judgments about frequency within the normal population and degree of aberrance in variability among IQ scores and indices within the profile. Judgments were measured on a seven-point Likert-scale. The questions for the two dependent variables were: "How frequently does the level of variability displayed in the profile below occur in the normal population?" and "In regards to intra-test variability, how would you rate the level of aberrance, if any, in the profile below?" Corresponding Likert-scale anchors for the first question were: '1' – "less than 10% of the time," '4' – "around 50% of the time," and '7' – "greater than 90% of the time." Corresponding Likert-scale anchors for the second question were: '1' – "strongly suggests normality," '4' – "neutral, no more likely to indicate normality or abnormality," and '7' – "strongly suggests abnormality" (see Appendix A). Given the small sample for this pilot study, analysis consisted of descriptive and non-parametric (Wilcoxon-Mann-Whitney) statistics.

Results showed that participants in the SS group judged the inter-test variability within the profile to be more common (M = 5.8, SD = 0.8) and less aberrant (M = 2.6, SD = 1.8) than participants in the PES group (M and SD = 4.0, 1.8; 4.7, 1.2, respectively). Despite the limited sample size, the analysis yielded significant differences for both variables (Z = 1.79, p = .04; Z = -1.67, p = .05, respectively). As the profiles were mathematically constant and the only difference was graphical display, the NIA between indices and IQ scores likely explained the results. Thus, the

pilot study provided initial evidence that NIA influences interpretations. Of note, two variables were simultaneously altered within the graphical display (i.e., visual scatter and metric selection). Therefore, whether both variables contributed to the outcome and their relative contributions could not be determined.

Hypotheses

The pilot study justified further investigation on the potential impact of NIA on test data. Differences in the pilot study were presumably due to alterations in the appearance of variability. Visual scatter[2] appears greater when results are plotted in PES versus SS. The pilot study also provided preliminary evidence that mode of graphic display may worsen certain interpretive practices, such as analysis of scatter – a practice that is already questionable in and of itself. Based on the results from the pilot study, the present study examined the following hypotheses:

H1: Participants would judge scatter to be associated with neuropsychological dysfunction and would underestimate the level of scatter found in healthy individuals.

H2: Greater versus lesser visual scatter (i.e., profiles plotted as percentiles with equal- vs. unequal-sized units) would lead participants to judge the test results as more unusual.

H3 (exploratory): Metric selection would influence judgments regarding scatter even when mathematical properties and visual scatter of test data are held constant.

[2] *Visual scatter* can be differentiated from *analysis of scatter*. Visual scatter refers to the visual appearance of data when plotted graphically, not to mathematical properties of the data. Analysis of scatter refers to the interpretation of scatter, which may or may not rest in part or in whole on mathematical properties.

H4: Truncating visual scatter, by maintaining the overall dimensions of the graph but increasing the range of anchor points along the y-axis, would reduce participants' overinterpretations of test scatter

CHAPTER 3

METHODOLOGY

The main study extended the pilot study by making the visual stimuli more representative of common clinical neuropsychological practice, by graphing data at varying levels of scatter, and by exploring a corrective procedure (detailed below).

Participants

Participants (N = 193) were recruited from the NPSYCH Listserve, an e-mail discussion list devoted to practice and research in adult neuropsychology. It is one of the more active neuropsychology listserves (NAN, 2003). It currently has 2,745 subscribers and is only open to neuropsychologists and other related specialists and researchers (accessed from the listserve's website, www3.npsych.com). Almost all members are expected to have had specialized training in neuropsychological assessment. There were no exclusionary criteria based on demographic features. A brief description of the study with a link to the survey (hosted by www.qualtrics.com) was posted to the NPSYCH Listserve on four occasions. The Qualtrics program randomly provided participants with one graphical display.

Out of the 193 participants, some participants did not provide responses to all demographic variables and, on rare occasions, did not respond to one of the dependent variables. For example, only 171 participants responded to the question regarding gender, 174 participants responded to the first dependent variable, and 185 participants responded to the second dependent variable. Missing data will be addressed below.

This study followed American Psychological Association ethical guidelines and was approved by the University of Rhode Island Institutional Review Board on Human Subjects.

Demographic features of the sample are summarized in Table 1 (see Appendix B for the demographic questionnaire). The sample was split about evenly by gender and was predominately White (90.7%). Twenty participants (11.7%) reported their level of education as M.A/M.S. Predoctoral level participants were included in the current analysis, because subscription to the NPSYCH Listserve suggests some familiarity with neuropsychological assessment. Inclusion of predoctoral level participants introduces limitations to the generalizability of this study.

Procedure

The study used test results from a prototypical neuropsychological battery (described below). Two levels of scatter or intra-individual variability were used and was defined, as is common, by the range between an individual's highest and lowest scores across the test battery. At each level of scatter, information was kept mathematically constant, and the only variation involved a NIA.

The levels of scatter were based on Schretlen et al.'s (2003) results. Schretlen et al.'s work had various positive design features (e.g., a more substantial normative sample than many studies of this type) and seemed to provide a strong template and basis for the current research. Schretlen and colleagues studied 197 healthy adults, age 20-92 with a mean age of about 55 years and a mean education of about 14 years. Each participant completed a neuropsychological battery of 15 tests that resulted in 32 measures or scores. Schretlen et al.'s study revealed substantial intra-individual

variability in the performance of presumably healthy, normal adults. For example, only 2% of the sample obtained a range of scatter of less than two standard deviations (*SD*), whereas 65% demonstrated a range of at least three *SD* and 20% a range of at least four *SD*. The mean level of intra-individual variability was about 3.4 standard deviations (*SD* = 0.8). Schretlen et al.'s findings are congruent with decades-old analyses showing large variability in individuals' subtest scores across intelligence tests and on groups of tests included within neuropsychological batteries (for reviews of the literature, see Schretlen & Sullivan, 2013; Binder et al., 2009; Brooks et al., 2009; and Mitrushina et al., 2005).

For purposes of the present study, Schretlen et al.'s 2003 results were used as guides for estimating where levels of intra-individual variation fell relative to the general population (e.g., was it lower than usual) and for appraising the accuracy of respondent's assumptions about expected levels of intra-individual variability. Schretlen and colleagues also kindly agreed to provide data from their research on healthy individuals, which was to be used to assist in the design of the prototypical protocols. The design of these materials was already well underway, based on the information contained in the published study, before the underlying Schretlen data were obtained.

Rather than providing data from the 2003 study, Schretlen et al. provided an updated and more extensive data set (Schretlen, Testa, Winicki, Pearlson, & Gordon, 2008). For example, the size of the sample was increased by over 50% (*N* = 327), and data were collected from an expanded and somewhat altered set of tests. Furthermore,

although not realized by the present author at the time, results for mean level of intra-individual variation were revised in an upward direction (M = 3.9; SD = 0.7).

The ultimate consequences of these changes were that levels of scatter for the prototypical protocols, which were selected to fall at the 10th and 50th percentiles based on Schretlen et al's 2003 study, fell at the 2nd and 25th percentiles, respectively, according to the updated and expanded data base. The prototypical protocols had already been selected and sent to participants before these changes in normative values were uncovered, but, by sheer luck, the difference did not undermine data collection or analysis, as will be further explained below. In short, even under the original set of normative assumptions based on the 2003 data, results showed that respondents markedly underestimated normal levels of scatter, and as the updated data showed even greater levels of intra-individual variation among normal individuals, this finding was not negated but rather shown to be even more extreme.

The data base Schretlen and colleagues provided included demographic variables and test performances on 45 overall measures. Thirty-two of these 45 overall measures were selected to create the prototypical profiles[3]. Within the current study, the 32 measures assess: memory (12), attention (6), executive function (4), language (3), visual-spatial (5), and motor (2) abilities. As noted, profiles were selected to be prototypical for levels of scatter at 2.4 *SD* and 3.4 *SD*, approximately the 10th and 50th percentiles (although as described, the figures fell at the 2nd and 25th percentiles,

[3] Here again, the original design for the present study was based on Schretlen et' al's 2003 data, which contained 32 test scores. Although the updated data base provided scores on additional tests, 32 measures seemed fully sufficient for research purposes and hence was not expanded. Figures provided above on normative levels of scatter for the updated data base were based on these 32 measures or areas of assessment.

respectively, according to the updated data base). Due to a concern that neuropsychologists may frequently over-interpret small to normal scatter as unusually large and aberrant, a normatively, small level and average level of scatter was chosen.

Working from the data base Schretlen and colleagues provided, prototypical cases were selected by examining results on the 32 pertinent test scores or areas of measurement for level of scatter. Only one case had a discrepancy of 2.4 *SD* (T-score range: 36 – 60) and 19 cases had a discrepancy of 3.4 *SD*. Of the 19 cases, seven were missing at least one score. Of the remaining 12 cases, one case was selected at random (T-score range: 31 – 65). Therefore, one case was selected with a level of scatter at 2.4 *SD* and one with a level of scatter at 3.4 *SD* (or results that fell at the 10^{th} and 50^{th} percentile, respectively, according to Schretlen et al. [2003] but at the 2^{nd} and 25^{th} percentile, respectively, according to the updated data base) .

As the intended participants were neuropsychologists and likely had varying familiarity and opinions about the tests that Schretlen et al. (2003; 2008) studied, generic labels for the tests were used. For example, tests were labeled by cognitive domain (e.g., Executive Function), instead of using the actual names of tests/subtests.

To assess the impact of NIA on interpretations, this study utilized a between-groups design with four independent variables and two dependent variables. Participants were provided with one out of seven possible profiles (detailed below). Each participant was then asked to judge how often the level of scatter displayed in the profile occurs in the normal population. Unrelated to the graphical display, participants were also asked whether scatter is associated with neuropsychological

dysfunction. If participants responded affirmatively, they were asked to approximate a dividing point that distinguishes between normal and abnormal levels of scatter.

Independent Variables: Scatter, Interval Spacing, Metric, and Corrective Procedure

The first independent variable was inter-test scatter. Inter-test scatter was set at 2.4 *SD* and 3.4 *SD*, which, as noted, turned out to approximate the 2nd and 25th percentile, respectively, in the data Schretlen provided. It seemed worth examining whether susceptibility to NIA varies across differing levels of scatter. For example, it may be that at 3.4 *SD*, results would be mistakenly judged as aberrant even if the graphical display attenuates the appearance of scatter.

The second independent variable, interval spacing, had two conditions: percentiles with equal-sized units (PES) versus percentiles with unequal-sized units (≠PES) (see Figures 2 – 5). As mentioned, interval spacing plotted as ≠PES is proportionate to a normal distribution. Both profiles were plotted with the same metric (percentiles) but visual scatter was greater with the PES profile, and therefore, the only difference between the PES and ≠PES profile involved NIA of scatter. The third independent variable, metric selection, was crossed with the second independent variable and also had two conditions: T-score vs. ≠PES, as shown in figures 6 and 7. The profile plotted as ≠PES had the same visual scatter as the T-score profile, and therefore the only difference between the T-score and ≠PES was the denoting metric.

Finally, a corrective procedure to attenuate the potential impact from NIA was explored. This fourth independent variable was crossed with the T-score profile plotted at 3.4 SD. As noted earlier, when a NIA accentuates visual scatter (e.g., PES), it seems likely profiles will be misjudged as rarer or much rarer occurrences than is

truly the case. Such rarity might well be overinterpreted or misinterpreted, in turn, as indicative of pathology. Therefore, these profiles may cause considerable interpretive problems and create a pressing need for corrective procedures or interventions.

The intervention examined here extended the range of T-scores plotted along the y-axis (from the original plotted range of 20-80 to the extended range of 0-100), while keeping the vertical and horizontal length, or overall dimensions, of the graph constant. By keeping overall vertical and horizontal size constant but extending the range of plotted T-scores, the physical distance between higher and lower test scores is diminished or condensed (see Figure 8 in comparison to Figure 7).

Dependent Variables

Participants were initially provided with a brief vignette of a hypothetical client. Demographic variables regarding age, education, and gender of the hypothetical client were borrowed from the mean characteristics found in Schretlen et al. (2003). The instructions were as follows:

> A neuropsychological profile of a 55 year old, white, female patient with 14 years of education appears below. The comprehensive battery consists of 15 tests, which generate 32 scores that have been co-normed (see below for list of test domains). All scores were uniformly adjusted for age, gender, and education. Although the data provided are certainly less complete than would be typical in practice, they should be valuable in addressing the questions that follow.

There were two primary dependent variables: a) judgment about the frequency of scatter within the graphical display and b) perceived relationship between scatter

and neuropsychological dysfunction. The question for the first dependent variable was: “How frequently does the level of inter-test variability displayed in the profile occur in the normal population?” Responses were measured on a continuous scale ranging from 0% to 100%. For the second dependent variable, the initial question used a dichotomous format and was followed by a subsequent question dependent on the first response. This second dependent variable was unrelated to the graphical display, and it was intended to assess broader interpretive practices. For the initial question, participants were asked: “Is inter-test variability associated with neuropsychological dysfunction?” If the response was affirmative, the participant was then asked to complete a second question: “Taking all test scores into account and assuming that 15 co-normed tests with 32 measures have been administered, what is the approximate dividing point you use in distinguishing between normal and abnormal levels of variability?” Corresponding Likert-scale anchors for this question were the maximum discrepancy between the highest and lowest score measured in standard deviations ranging from less than 0.5 *SD* to greater than 6.0 *SD*. (see Appendix C).

The two dependent variables were treated as independent from one another. The first dependent variable was used to address the second through fourth hypotheses, which all involved the impact of NIA on interpretation. The second dependent variable was used to address the first hypothesis, which involved judgments regarding scatter and neuropsychological dysfunction.

Analysis

Based on the results of the pilot study, a medium to large effect size (d = .325) was anticipated. To achieve 80% statistical power, an overall sample of 245

participants (35 per cell) was needed. The overall design was segregated into three separate Analyses of Variances (ANOVAs) based on the groupings of the independent variables: interval spacing, 2X2; metric selection, 2X2; and corrective procedure, 1X1. Given the crossed nature of the design, there are seven cells overall. As described below, the actual analysis for the study utilized two one-way Mann-Whitney *U* tests (comparable to two one-way ANOVAs). Each analysis was performed at the two levels of scatter (i.e., 2.4 *SD* and 3.4 *SD*). An *a priori* power analysis (based on calculations using G*Power 3.1.3) was conducted on the PES vs. ≠PES groups at both levels of scatter to calculate an adequate cell size. Given the three separate ANOVAs planned for the study, a Bonferroni correction was employed to maintain an overall type I error rate of 5%. Thus, the α level was set at .017. Assuming the same parameters mentioned above (in particular the effect size), a *post hoc* power analysis on the obtained sample (N = 193) indicated 60% statistical power. However, statistical analytic techniques assumed in the power analysis were ultimately modified. The above procedure still served to guide the initial recruitment and study design.

The original plan anticipated a Multivariate Analysis of Variance (MANOVA) with subsequent follow up Analysis of Variance (ANOVA) and Tukey Tests as indicated. However, the overall sample size was suboptimal, and Levene's test indicated unequal variances (F = 2.198, p = .05). Therefore, non-parametric tests that are comparable to ANOVA and follow-up Tukey Tests were used; specifically, Kruskal-Wallis tests with follow-up one-tailed Mann-Whitney *U* tests. Mann-Whitney *U* tests were set *a priori* as one-tailed to account for the expected directionality in responses (i.e., responses were expected to skew toward normative characteristics

based on the level of scatter). The second dependent variable (relationship between scatter and neuropsychological dysfunction) was analyzed with descriptive statistics.

CHAPTER 4

FINDINGS

H1: Is Scatter Associated with Neuropsychological Dysfunction?

It was initially hypothesized that participants would judge scatter to be related to neuropsychological dysfunction. It was further hypothesized that those participants who affirmed a relationship would underestimate normal level of scatter in healthy individuals. Participants who did not affirm a relationship obviously could not make a judgment regarding level of scatter differentiating normal and abnormal performance. Results were analyzed by means of descriptive statistics.

Eight participants (4.1%) did not respond to the question asking whether inter-test variability is associated with neuropsychological dysfunction. Of the 185 who did respond, 64.3% (n = 119) indicated that scatter is associated with neuropsychological dysfunction. Among those responding affirmatively, 72% (n = 86) indicated a dividing point for abnormal levels of scatter at 1.5 *SD* to 2.0 *SD*. Slightly less than 2% of respondents (only two) indicated a cutoff at 4.0 *SD*, and none indicated a cutoff below 1.0 *SD*. Table 2 provides cumulative percentages for level of scatter judged to distinguish between normal and abnormal performance. Many participants dramatically underestimated a cutoff for determining abnormal levels of intra-individual variability or scatter not only when compared to Schretlen's actual results (e.g., 3.0 *SD* and 3.9 *SD* approximated the 10^{th} and 50^{th} percentiles) but also when

compared to numerous other studies on the matter that have yielded similar outcomes (see Binder et al., 2009)

Main Effects for Non-Informative Alterations at Both Levels of Scatter

To evaluate the main effects for NIA, Kruskal-Wallis tests were conducted to examine differences among the three conditions with 2.4 *SD* of scatter (PES, ≠PES, and T-score) and the four conditions with 3.4 *SD* of scatter (PES, ≠PES, T-score, and T-score with intervention). The outcome was significant at 2.4 *SD* of scatter, χ^2 (2, N = 72) = 6.35, $p < .05$; and at 3.4 *SD* of scatter, χ^2 (3, N = 102) = 28.04, $p < .001$. Follow-up Mann-Whitney *U* tests were conducted to evaluate pairwise differences among the groups at each level of scatter. Descriptive statistics are summarized in Table 3. Respondents demonstrated a wide range of judgments regarding perceived frequency when evaluating both levels of scatter. That is, at both levels of scatter, participants' judgments of frequency ranged from zero to 95%.

H2: Interval Spacing (PES vs. ≠PES)

It was hypothesized that greater versus lesser visual scatter (i.e., profiles plotted as PES vs. ≠PES) would lead participants to judge the test results as rarer. At 2.4 *SD* (Figure 2 vs. 3), a one-tailed Mann-Whitney *U* test revealed no significant difference between the PES (*Mdn* = 35) and ≠PES (*Mdn* = 35) conditions, U = 309, Z = -.068, p = .48. The effect size[4] (r = .01) was negligible (Cohen, 1992). Similarly, at 3.4 *SD* (Figure 4 vs. 5), a one-tailed Mann-Whitney *U* test revealed no significant difference between the PES (*Mdn* = 22) and ≠PES (*Mdn* = 25) conditions, U = 377.5, Z = -.206, p = .42, again with a negligible effect size (r = .03). Contrary to the

[4] The non-parametric effect size was approximated using the following equation: $r = Z/\sqrt{N}$ (Rosenthal, 1991).

hypothesis, no difference in judgment was found between test data plotted as percentiles in a manner congruent or incongruent with the underlying mathematical properties (i.e., as ≠PES versus PES, respectively).

H3: Metric Selection (≠PES vs. T-score)

The exploratory hypothesis that metric selection would influence judgments regarding scatter was supported. Even when mathematically equivalent data was plotted identically, simple change in the designated metric altered judgments. At 2.4 *SD* (Figure 3 vs. 6), a one-tailed Mann-Whitney *U* test revealed a significant difference between the ≠PES (*Mdn* = 35) and T-score (*Mdn* = 65.5) conditions, $U = 172.5$, $Z = -2.187$, $p < .05$, with a medium effect size ($r = .32$). Similarly at 3.4 *SD* (Figure 5 vs. 7), a one-tailed Mann-Whitney *U* test revealed a significant difference between the ≠PES (*Mdn* = 25) and T-score (*Mdn* = 40) conditions, $U = 191.5$, $Z = -2.345$, $p < .001$, again with a medium effect size ($r = .33$). Thus, participants judged equivalent data and plotting of results as more common when designated as T-scores versus percentiles across both the 2.4 *SD* and 3.4 *SD* conditions.

H4: Corrective Procedure (T-score vs. T-score with Intervention)

Based on the first hypothesis that greater visual scatter would lead participants to judge the test results as rarer, a corrective procedure was implemented. It was hypothesized that reducing or truncating the physical space between high and low test scores would improve judgments about the frequency of scatter, or bring them in closer alignment with research findings. The corrective procedure maintained the overall dimensions of the graph but increased the range of anchor points along the y-axis. This analysis was conducted only at a scatter of 3.4 *SD* (Figure 7 vs. 8). A one-

tailed Mann-Whitney *U* test revealed a significant difference between the T-score (*Mdn* = 40) and T-score with intervention (*Mdn* = 75) conditions, *U* = 112.5, *Z* = -3.335, *p* < .001. The effect size (*r* = .49) was large. Consistent with the hypothesis, a visual representation that truncated the appearance of scatter altered judgment about the frequency of occurrence, and it did so in a favorable direction.

CHAPTER 5

CONCLUSION

This study examined: (1) judgments regarding scatter across a neuropsychological battery, (2) whether NIAs impact judgments regarding scatter, and (3) whether truncating the visual presentation of scatter attenuates or corrects misjudgments about the normality of scatter. When NIA impacts interpretation, it likely degrades the accuracy of psychological and neuropsychological evaluation. Additionally, when NIA exerts an impact, clinicians might underutilize diagnostic, or truly useful, information. Interpretive practices based on configural relationships may be particularly vulnerable to influence from NIA, given emphasis on patterns and interrelations among tests scores, which can look very different depending on variations in visual presentation. Although research has recognized limitations in scatter analysis for over half a century (Schofield, 1952), the appraisal of test scatter remains one of the most common approaches to evaluation of cognitive function and brain disorders (Lezak et al. 2012). If or when NIA degrade the accuracy of certain interpretive practices, corrective procedures for graphically displaying data could prove highly beneficial.

Neuropsychologists' Perceptions of Normal Scatter

In this study, a substantial proportion of neuropsychologists (64.3%) endorsed the value of scatter for identifying neuropsychological dysfunction. Those endorsing the value of scatter were then asked to specify a cutoff for abnormally high levels

under the assumptions set forth in the research materials, which involved 32 co-normed tests or test scores. Although specified cutoffs varied all the way from 1.0 *SD* to 4.0 *SD*, nearly every respondent underestimated normal levels of scatter; many by a large margin. For example, 72.6% indicated a cutoff between 1.5 *SD* and 2 *SD*, levels well below those expected for *normal individuals* and very often exceeded by such groups. The vast majority of normal individuals in the Schretlen data base (over 99%) exceeded the range of 2 *SD*.

Although the Schretlen data provide a single source of information on scatter and are not definitive, the level of scatter found in that work is consistent with a considerable body of literature on the topic (Binder et al., 2009; Brooks et al., 2009). Consider further that studies involving even a single general measure with about 10 or so subtests, such as the Wechsler Intelligence scales, demonstrate levels of scatter among normal groups that equal or exceed the cutoff levels that many respondents in the current study identified under the assumption that about triple the number of measures were used. For example, the 11 primary subtests from the WAIS-III and the 10 primary subtests from the WAIS-IV both have a mean of about 2.2 *SD* between the highest and lowest scores (Wechsler, 1997a; 2008). It is also a mathematical truism that increasing the number of tests or subtests within a neuropsychological battery that already includes such an intelligence test will produce a level of scatter that must at least equal, and will often exceed, the level of scatter produced by the intelligence test alone (Binder, Iverson, & Brooks, 2009).

Furthermore, neuropsychological batteries are often comprised of various measures that are *not* co-normed, which is likely to accentuate scatter. Variability

between test scores and measures may also be magnified by various artifacts, such as scoring errors (Allard & Faust, 2000; Simons, Goddard, & Patton, 2002) and inadequate normative selection (Brooks, Strauss, Sherman, Iverson, & Slick, 2009). Taking all of these considerations together, a critical implication of the current results is that common interpretive practices, which both emphasize scatter analysis and grossly underestimate normative levels, may well lead to the overidentification of pathology or brain dysfunction, a potentially serious error.

The already problematic practice of overinterpreting scatter may be worsened by NIA because the visual distance between graphically displayed test scores is partly determined, and may be subsequently altered, by interval spacing, metric selection, and graphical dimensions. This study and prior research of Schretlen et al. (2003) demonstrated that psychologists frequently underestimate normal levels of scatter. Therefore, graphical presentation that accentuates differences between scores may compound these interpretive problems, a possibility the current study partly supports. Identification of such issues provides a platform for exploring possible interventions, such as truncating visual representations of scatter to decrease perceived rarity.

Non-Informative Alterations

This study explored two primary NIAs: 1) interval spacing of percentiles, and 2) metric selection with identical visual scatter. The influence (or lack thereof) from these alterations on judgments was similar for each variable at both levels of scatter. Contrary to the hypothesized result, interval spacing of percentiles did not result in significant differences at either level of scatter. However, metric selection did produce

significant differences at both levels of scatter, with scatter represented in T-scores judged to be more common (or less aberrant) than percentiles.

Interval Spacing (Percentiles with Equal- vs. Unequal-Sized Units)

As mentioned earlier, one limitation of percentiles is the lack of equal interval measurement, resulting in disparity in the relationships between test scores. It was hypothesized that scatter plotted in percentiles at equal-sized units would be judged as rarer than scatter plotted in unequal-sized units, because in the former case the mathematical distortion artificially increases physical distance between test scores. When percentiles are plotted in equal-sized units, even scores that do not fall very far from the middle of the bell curve (e.g., a score 1 standard deviation below the mean) are nevertheless pulled towards the endpoints of a graphical display. Record forms (Wechsler, 1997a), computer-based test interpretive programs (The Psychological Corporation, 2002b), and authoritative texts in neuropsychology (Spreen & Strauss, 1998) have graphically displayed percentiles as equal-sized units. However, in the present study, no difference was found at either level of scatter for equal- versus unequal-sized units

Metric Selection (Percentiles vs. T-scores)

Unlike the non-significant results obtained when altering the representation of percentiles, simply labeling identical visuals plots as percentiles versus T-scores led respondents to judge test variability as less common or more aberrant (median range for percentiles = 22% – 35%; median range for T-scores = 40% – 75%). It is possible that the mere magnitude of the percentile *number* influences judgments of the profile's rarity. For example, a maximum discrepancy represented in T-scores of 30 and 70 is

equivalent to a percentile of approximately 2 and 98. Perhaps the magnitude of the difference between listed numbers can be more salient than the magnitude of the visual discrepancy. These findings and conclusions are preliminary and merit further research. If, however, the mere number itself, and not the correct meaning of the number (based on both the number and its true value given the respective metric) exerts an impact, and perhaps a decided impact, on clinical judgment, it is a cause for serious concern.

When the criterion level of scatter was set at an unusually low level (the 2nd percentile or 2.4 *SD*), judgments about frequency or rarity based on T-scores were more accurate than judgments based on percentiles. However, even these appraisals based on T-scores were still markedly inaccurate (*Mdn* = 35%). Similarly, a criterion level of scatter set at the 25th percentile (i.e., 3.4 *SD*) was appraised more accurately when represented as T-scores versus percentiles (*Mdn* = 40%). Two judgmental problems may be occurring: a) misappraisal of normal scatter (as described above under "neuropsychologists' perceptions of normal scatter"), and b) inconsistent interpretability of and transposition between varying metrics (i.e., equivalent data are treated as unequal due to how the metric is conceptualized). In regards to the latter, this study provides evidence that certain metrics are either systematically misinterpreted or poorly understood in their relationship to a normal distribution. As a result, percentiles and T-scores that represent mathematically identical information may be interpreted differently.

Such mixed judgmental tendencies are potentially problematic. Variations in metric selection occur across both test record forms and computer-based test

interpretive programs. To record and interpret data, clinical neuropsychologists frequently use multiple mediums that present data in different metrics. Depending on metric selection and its relation to the susceptibility of particular interpretive practices, even a comprehensive neuropsychological assessment system that provides a uniform metric across all measures within a battery may be prone to NIA. It is unclear whether certain metrics compound error into otherwise problematic judgment practices (e.g., overinterpretation of scatter). The non-trivial impact of NIA in assessment is likely pervasive within clinical neuropsychology, and thus attempts to ameliorate the influence seems warranted.

Corrective Procedure (Truncated Visual Scatter)

It was hypothesized that decreasing physical spacing of visual scatter across higher and lower tests scores would lead respondents to judge profiles as more common or less indicative of pathology. This outcome did not occur when data were plotted as PES vs. ≠PES, and under both conditions many respondents underestimated the commonality or normality of scatter.

Perhaps somewhat paradoxically, however, the corrective procedure produced encouraging results. When data were held mathematically constant but visual scatter was reduced by extending the anchor points of the y-axis (0 – 100), judgments were altered in the desired direction. With a normal distribution, extreme T-scores represented within the corrective procedure become nearly unattainable (i.e., a T-score of zero represents a score worse than at least one in three million). However, such a truncated display demonstrated feasibility in diminishing the role of NIA, which may improve clinical judgments. Consequently, it may be possible, for example, to

attenuate what appears to be marked tendencies to overinterpret normal levels of scatter in healthy individuals.

Limitations

One limitation of this study was the suboptimal sample size. The study aimed to recruit 245 participants (i.e., 35 participants per cell for the first dependent variable) but fell about 70 participants short (obtained *n*'s = 22 – 30 participants per cell). Decreased sample size and smaller than expected effect size (a priori prediction: medium – large; actual effect size: medium) decreased the study's overall statistical power. Further, the non-normal distribution necessitated the use of non-parametric statistical methods. A principle limitation of non-parametric analysis is that the underlying distribution is unknown, which restricts the ability to generalize beyond the data. Therefore, results from the current study should be interpreted and generalized with caution. However, the results do suggest that NIA can have a significant impact on judgment, a potential phenomenon meriting further research.

The prototypical profiles used within the study were based on an updated and expanded data base from Schretlen et al.'s ABC study (2003). The normative characteristics of scatter from the updated data base are not identical to this study's original criterion characteristics (i.e., descriptive data on scatter from Schretlen et al.'s ABC study, 2003). Perhaps of primary importance, the level of scatter in the updated data base (3.9 *SD*) exceeds the level Schretlen et al. (2003) obtained (3.4 *SD*). The differences in the data bases simply suggest that problems observed in the current study in the overappraisal of scatter are more extreme than Schretlen et al.'s initial or earlier figures from 2003 might indicate. About 80% of the participants in the current

study set a cut-off for scatter at less than or equal to 2 SD, and whether one depends on Schretlen's earlier or later data or various other studies on the same topic it is highly probable that such levels are very common among normal individuals completing a neuropsychological battery.

Another study limitation involves the restricted data provided to participants. In standard clinical practice, a neuropsychologist will likely have access to detailed records, interview data, and other corroborating information, all of which might provide useful information. Efforts were taken to provide basic demographic information and test data that would be sufficient to answer the interpretive question. Participants may have preferred to have more detailed information regarding the hypothetical patient or specifics about actual measures. However, decades of research suggests that clinicians reach more accurate conclusions overall if they disregard interview results and base their interpretations on test results alone (Faust & Ahern, 2012). Participants also viewed the graphical display on their personal or business computer. It is possible that the size and horizontal to vertical ratio of their monitors might have resulted in subtle distortions in the dimensions of the graphical display. All data points and anchors along the y-axis would have maintained their proportional relations, but it still possible that subtle visual discrepancies could have been present and influential.

Lastly, participants were asked a dichotomous question regarding whether scatter is associated with neuropsychological dysfunction. Participants were given the option to comment on their response, and many individuals indicated that the question is dependent upon context. This criticism is legitimate, in particular because different

disorders may lessen or increase scatter or leave it unaffected. However, most respondents (about 64%) affirmed a relationship between scatter and neuropsychological dysfunction, and almost three-quarters of the latter (about 72%) grossly underestimated normal levels of scatter as falling between 1.5 to 2.0 *SD*. Hence, concerns expressed in this thesis about overinterpretation of scatter seem warranted.

In summary, participants in this study who were primarily licensed, clinical neuropsychologists[5] misperceived normal levels of scatter as rare or aberrant when the criterion for normal scatter was based on the data Schretlen provided (i.e., scatter with a maximum discrepancy of $M = 3.9$, $SD = 0.7$). This poses a problem because normal level of scatter may frequently be perceived as abnormal and lead to overpathologizing. The influence of NIA was mixed. Surprisingly, percentiles represented at equal- and unequal-sized units did not alter judgments related to scatter. In contrast, metric selection produced a significant difference, with data presented as T-scores judged as more common than data presented as percentiles for overall levels of scatter at both 2.4 *SD* and 3.4 *SD*. An intervention that truncated visual scatter improved judgmental accuracy. That is to say, even when keeping data mathematically equal, truncating visual scatter lead to more appropriate judgments about the commonality of the outcome. This study provides suggestive evidence for conceptual problems that may be pervasive in the field regarding analysis of scatter and inconsistent interpretations across metrics. Future research should determine whether the findings can be replicated, advance the design of interventions as needed, and

[5] Graduate students trained in neuropsychological assessment made up 11.7% of the sample.

assist in developing evidence-based standards for representing graphical displays that diminish the influence from NIA and shape judgments in normative directions.

Table 1: Demographic Features

		n	Frequency*
Gender		171	
	Male	85	49.7%
	Female	86	50.3%
	no response	22	--
Ethnicity		172	
	American Indian/Alaskan Native	1	0.6%
	Asian or Pacific Islander	2	1.2%
	Hispanic/Latino	5	2.9%
	Caucasian/White	156	90.7%
	Bi-Racial	3	1.7%
	Other	2	1.2%
	Choose not to disclose	3	1.7%
	no response	21	--
Highest Degree		171	
	M.A/M.S.	20	11.7%
	Ph.D.	115	67.3%
	Psy.D.	31	18.1%
	Ed.D.	4	2.3%
	Other	1**	0.6%
	no response	22	--
Currently Licensed		170	
	Yes	151	88.8%
	No	19	11.2%
	no response	23	--
Board Certification		166	
	Clinical Neuropsychology	47	28.3%
	Other	21	12.7%
	None	98	59.0%
	no response	27	--
Forensic Involvement		174	
	Yes	73	42.0%
	No	101	58.0%
	no response	19	--

* Missing data were excluded when calculating overall percentages.
** One participant responded "other" and provided a text response of "post doctorate."

Table 2: Level of Inter-test Scatter Judged to Distinguish Between Normal and Abnormal Performance

Scatter	Cumulative Percentage
>5.5	0.0
5.0	0.0
4.5	0.0
4.0	1.9
3.5	5.6
3.0	14.0
2.5	20.6
2.0	57.9
1.5	92.5
1.0	100.0
0.5	100.0
0.0	100.0
Mean	1.96
SD	0.65
Median	2.0

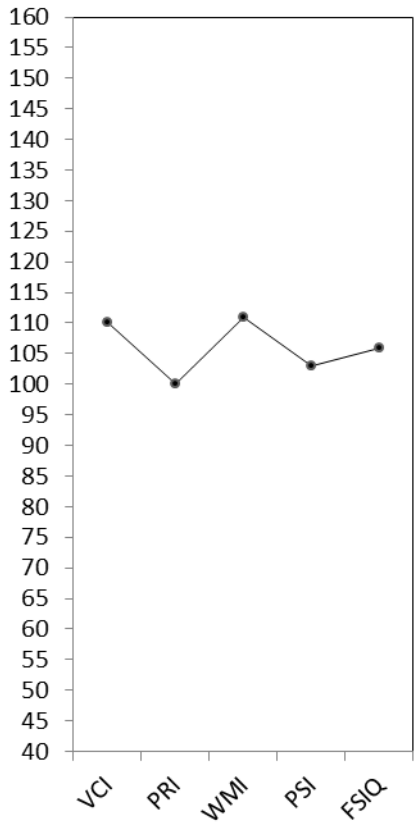
Note. Based on the results from participants who affirmed a relationship between scatter and neuropsychological dysfunction (i.e., 119 participants; 64.3%).

Table 3: Descriptive statistics by condition

	N	*M* (*SD*)	Median	Min.	Max
2.4 *SD*					
PES	25	41.6 (29.9)	35.0	0	92
≠PES	25	42.8 (29.9)	35.0	1	90
T-Score	22	61.0 (25.5)	65.5	5	95
3.4 *SD*					
PES	30	33.0 (27.4)	22.0	1	87
≠PES	26	27.7 (21.5)	25.0	1	80
T-score	24	44.0 (25.3)	40.0	0	85
Intervention	22	69.0 (21.7)	75.0	0	95

Figure 1: Pilot Study: WAIS-IV Prototypically Normal Profiles

Plotted as SS



Plotted as PES

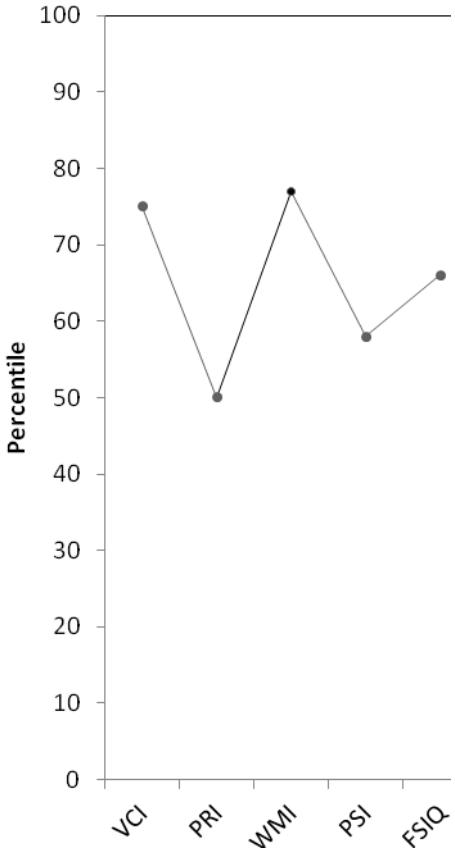


Figure 2: Percentile with Equal-Sized Units Set at 2.4 *SD*

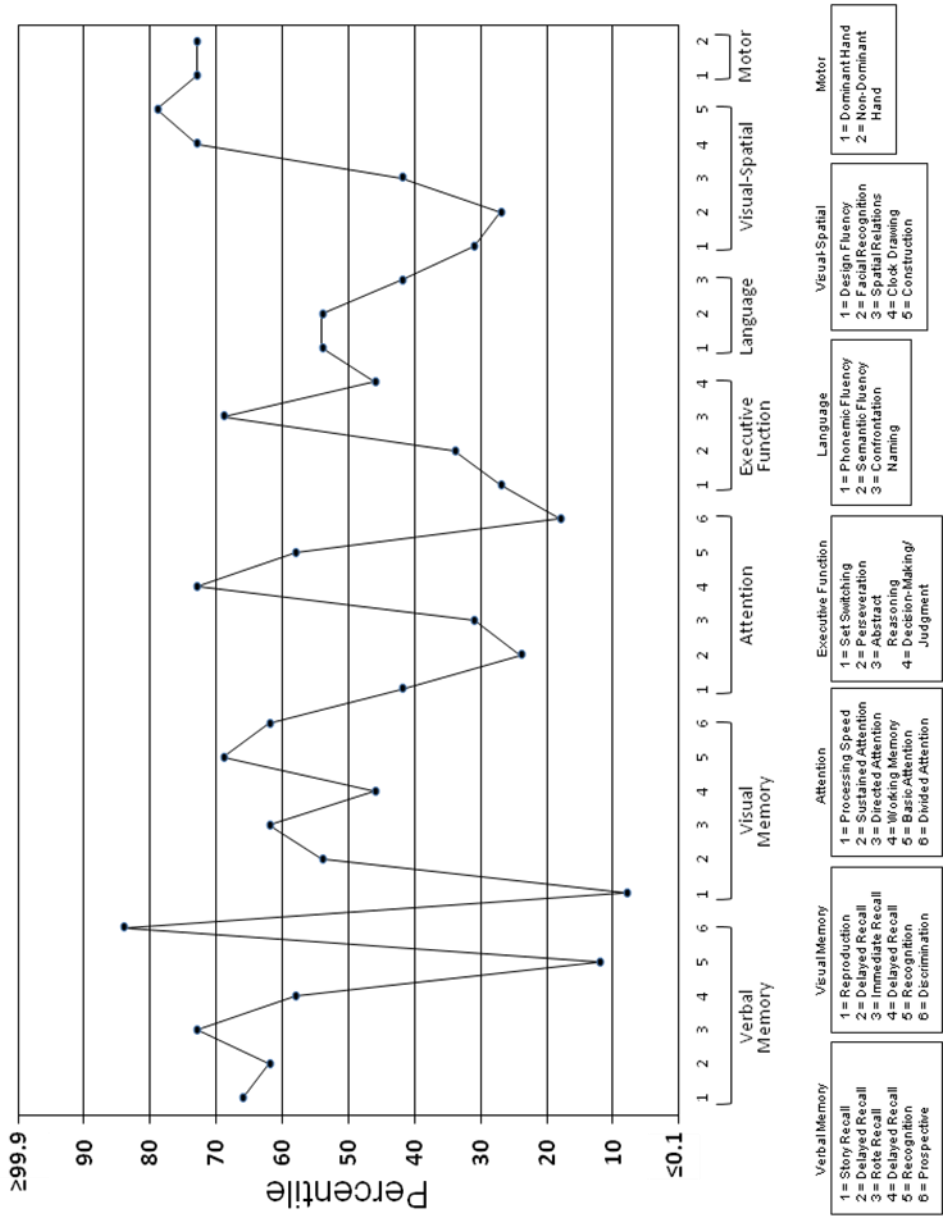


Figure 3: Percentile with Unequal-Sized Units Set at 2.4 *SD*

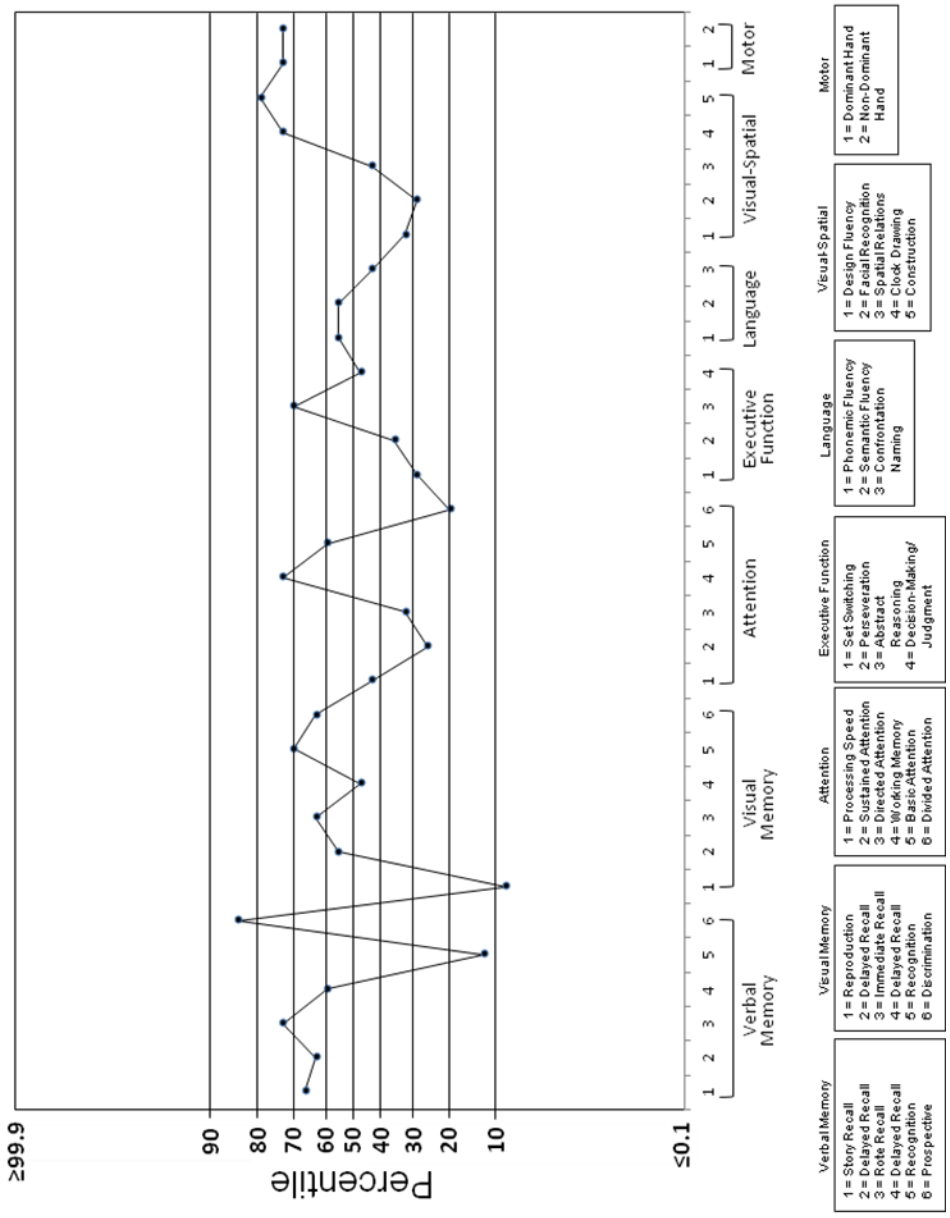


Figure 4: Percentile with Equal-Sized Units Set at 3.4 *SD*

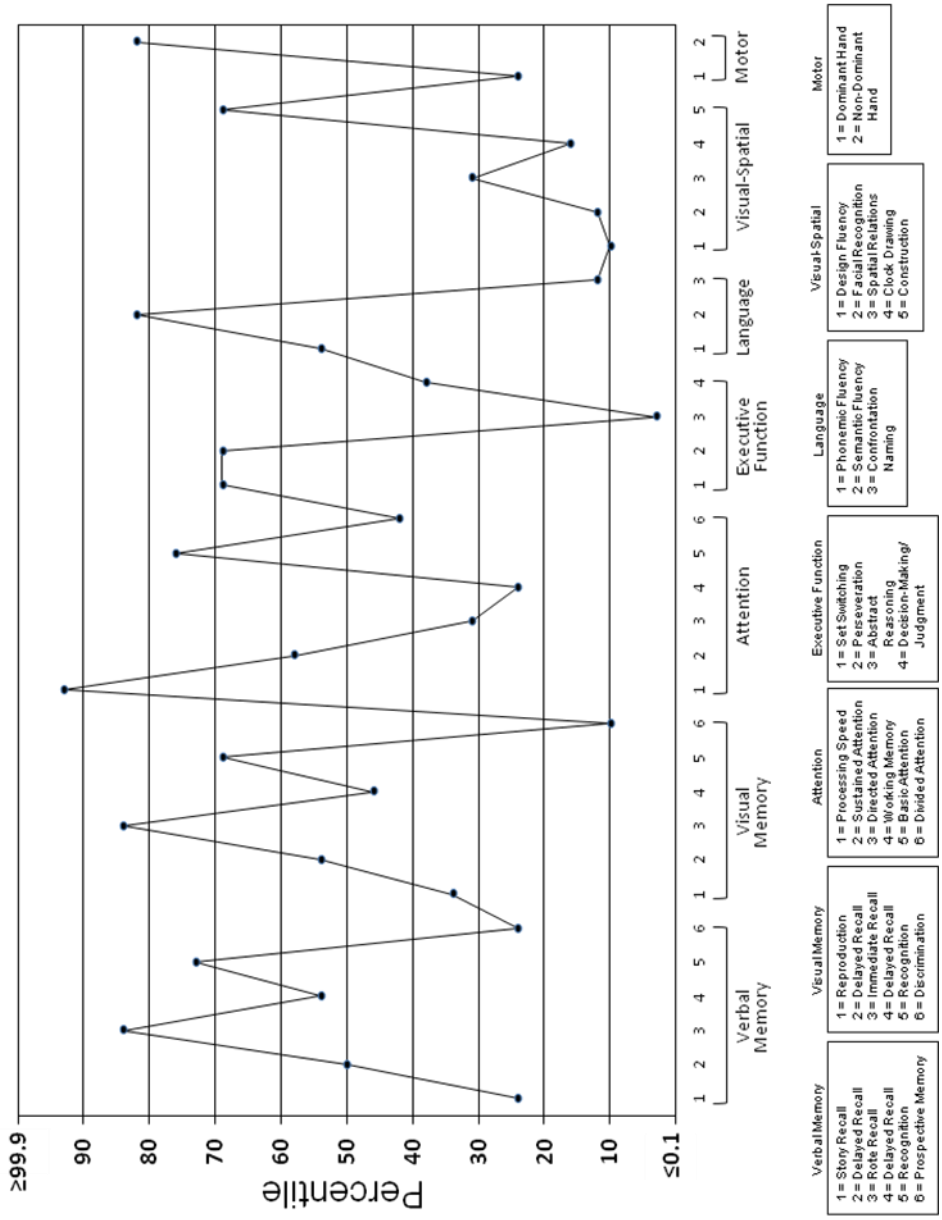


Figure 5: Percentile with Unequal-Sized Units Set at 3.4 *SD*

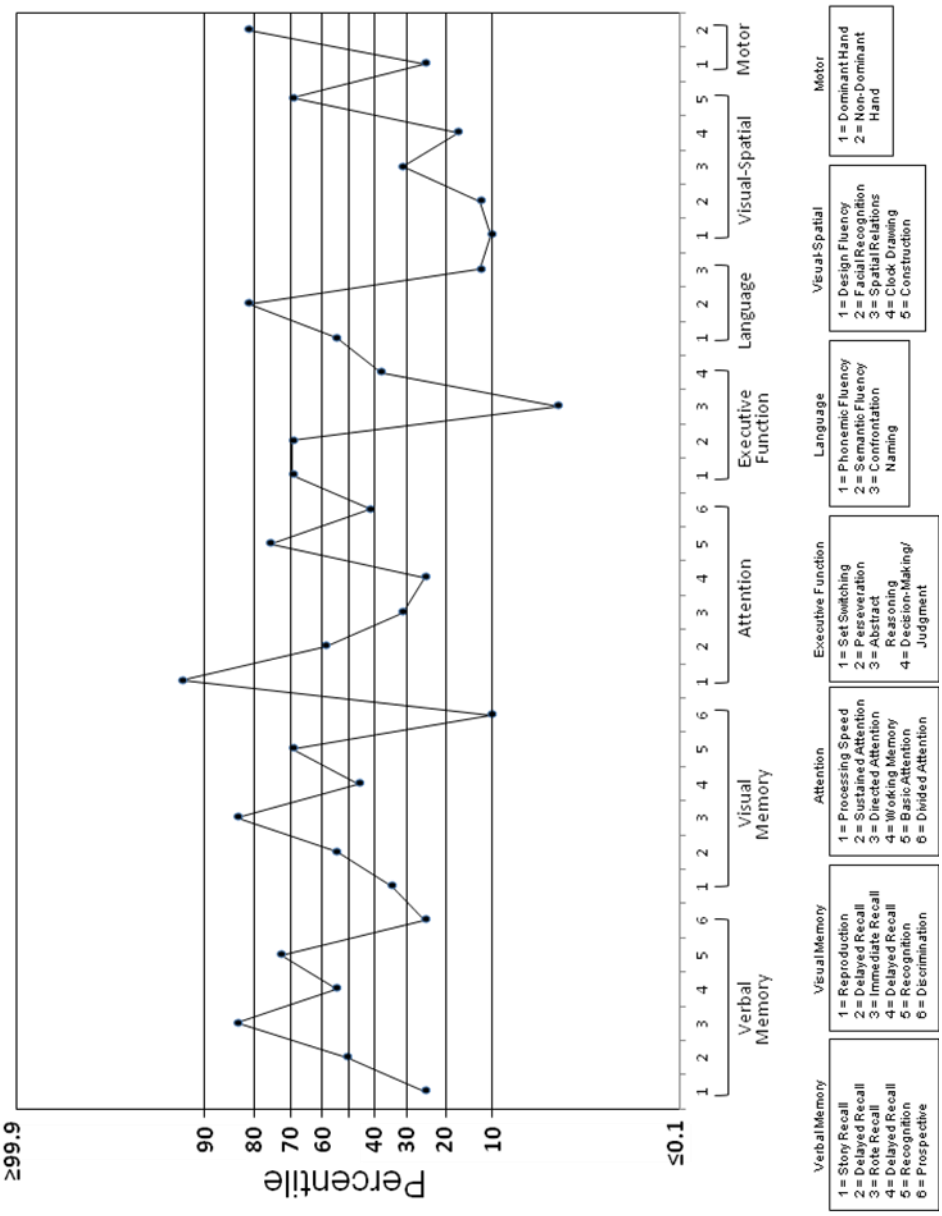


Figure 6: T-Score Set at 2.4 *SD*

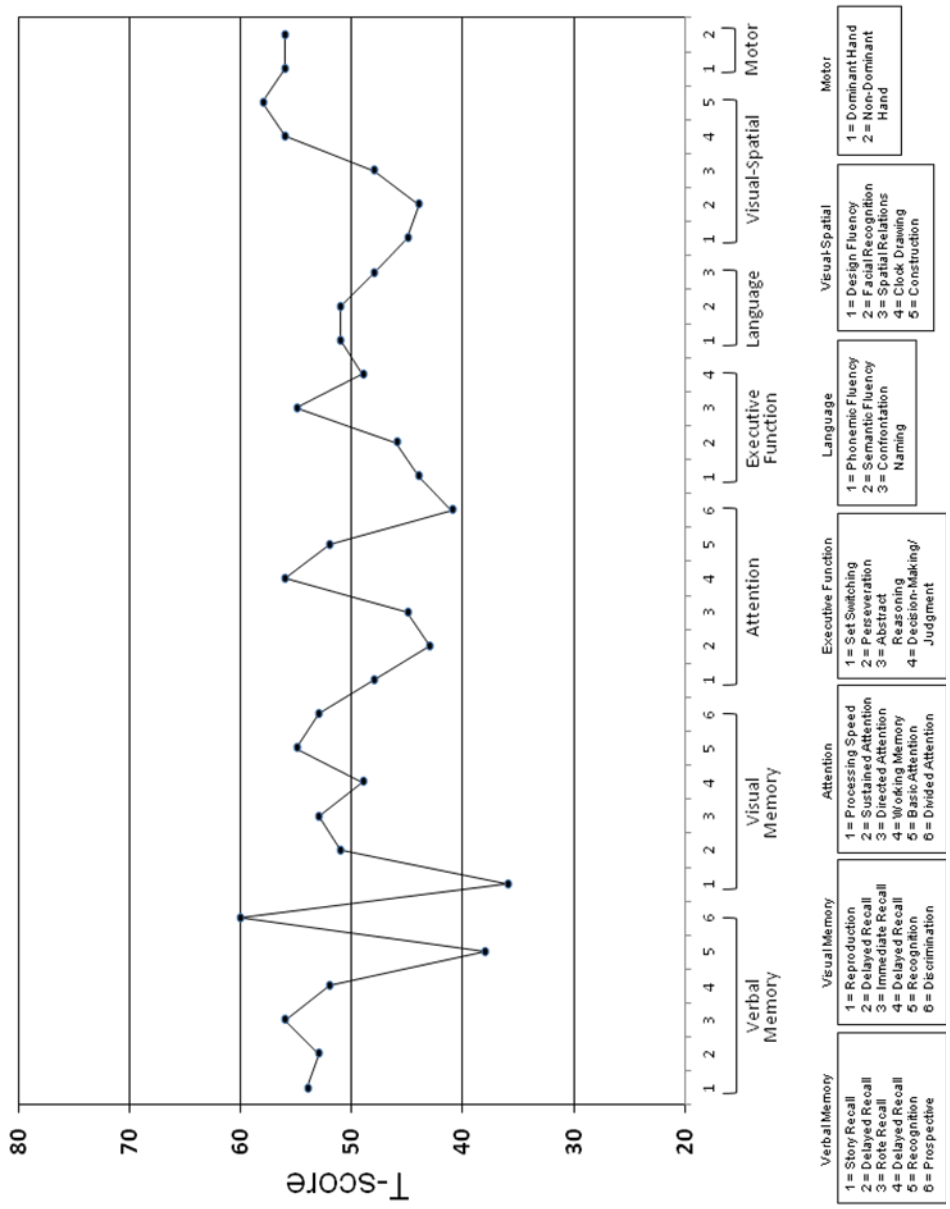


Figure 7: T-Score Set at 3.4 *SD*

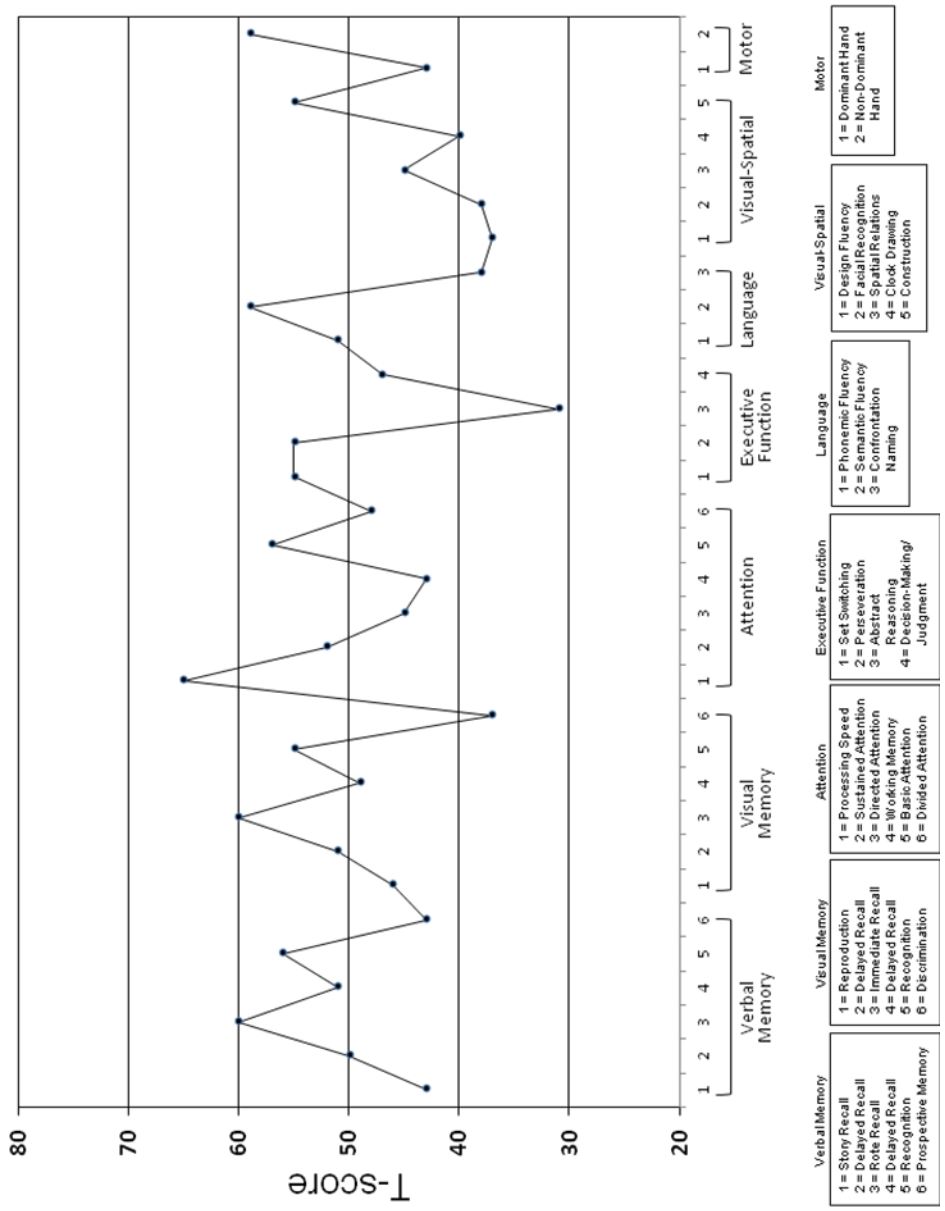
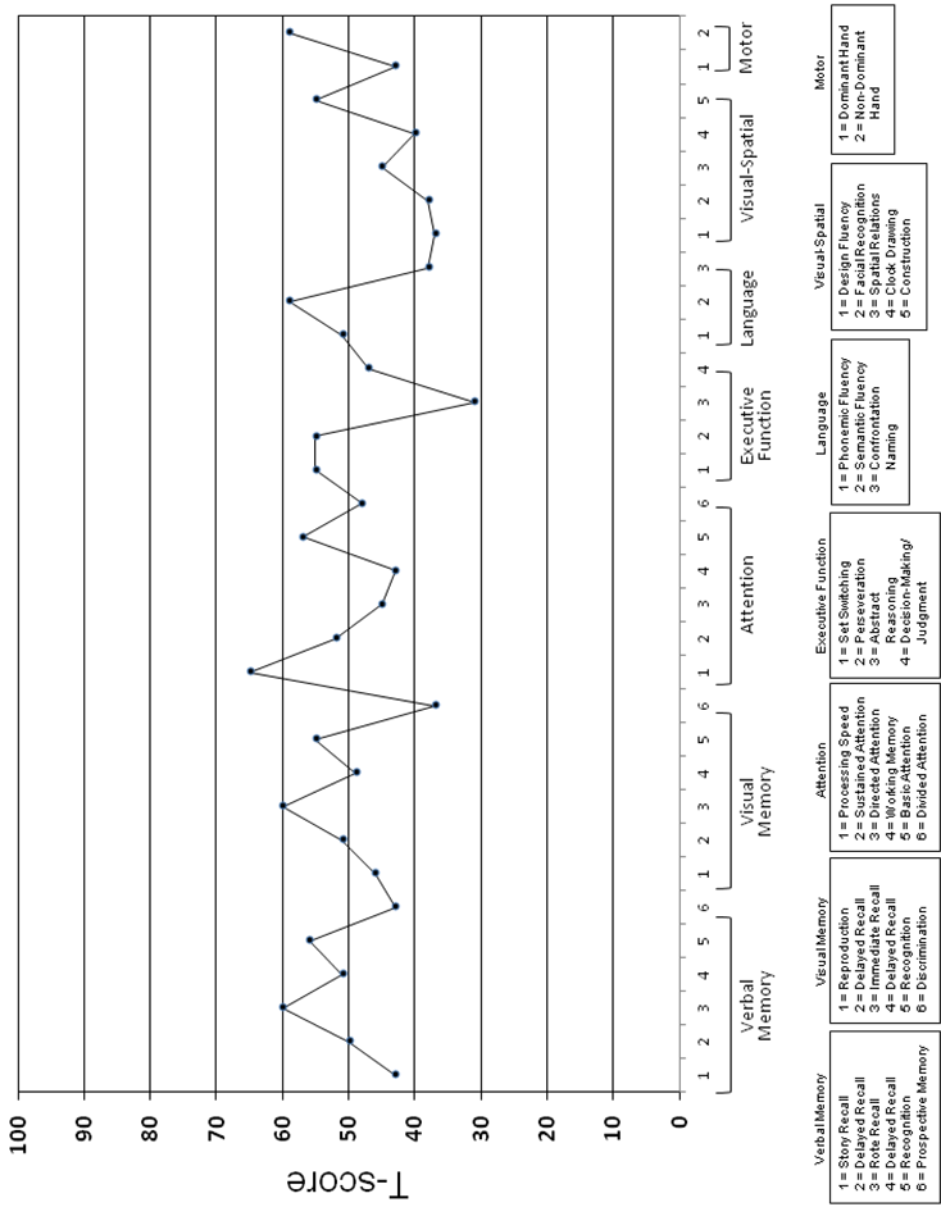


Figure 8: T-Score Set at 3.4 *SD* with Condensed Visual Scatter



APPENDIX A: Pilot Study: Dependent Variable Questions

Please respond to the questions below concerning the frequency and overall aberrance, if any, in the discrepancy between Wechsler IQ scores and Indexes. It is understood that in clinical practice the information provided is insufficient and other relevant information would be advantageous, e.g., referral question, demographic information and clinical history. Given the acknowledged limits of the information, please respond as best you can.

How frequently does the level of variability displayed in the profile below occur in the normal population?
(Please circle the corresponding number)

1	2	3	4	5	6	7
Less than 10% of the Time			Around 50% of the Time			Greater than 90% of the Time

In regards to intra-test variability, how would you rate the level of aberrance, if any, in the profile below?
(Please circle the corresponding number)

1	2	3	4	5	6	7
Strongly Suggests Normality			Neutral, No More Likely to Indicate Normality or Abnormality			Strongly Suggests Abnormality

APPENDIX B: Demographic Questionnaire

Please respond to the following questions concerning demographic information and professional practice.

1. **Gender:** Male Female

2. **Ethnicity:** African American/Black
   American Indian/Alaskan Native
   Asian or Pacific Islander
   Hispanic/Latino
   Caucasian/White
   Bi-racial
   Other
   Choose not to disclose

3. **Highest Degree:** B.A. M.A./M.S. Ph.D. Psy.D Ed.D Other

4. **Years since Highest Degree:** <5 5-10 11-20 >21

5. **Currently Licensed as a Psychologist:** Yes No

6. **Board Certification in Clinical Neuropsychology:** Yes No

7. **Board Certification in other specialty:** Yes No

8. **Over the last year, about what percentage of your time per week is spent on neuropsychological evaluations:**

   0% 1-25% 26-50% 51-75% 76-100%

9. **What percentage of your time is spent with the following populations::**

Children and Adolescents (≤18 years)	0%	1-25%	26-50%	51-75%	76-100%
Adults (19-65 years)	0%	1-25%	26-50%	51-75%	76-100%
Geriatric Adults (>65 years)	0%	1-25%	26-50%	51-75%	76-100%

10. **Are you involved in forensic evaluations:** Yes No

    If yes, over the last year, about what percentage of your time per week is spent on forensic evaluations:

    N/A 0% 1-25% 26-50% 51-75% 76-100%

11. **When available, how frequently do you use computer-based test interpretation programs when available:**

    0% 1-25% 26-50% 51-75% 76-100%

12. **What metric do you most commonly use during test interpretation:**

    Z-scores T-scores Wechsler SS Percentiles Grade Equivalent
    Other

13. **Which metric do you most prefer to use during test interpretation:**

    Z-scores T-scores Wechsler SS Percentiles Grade Equivalent
    Other

APPENDIX C: Main Study: Dependent Variable Questions

A neuropsychological profile of a 55 year old, white, female patient with 14 years of education appears below. The comprehensive battery consists of 15 tests, which generate 32 scores that have been co-normed (see below for list of test domains). All scores were uniformly adjusted for age, gender, and education. Although the data provided are certainly less complete than would be typical in practice, they should be valuable in addressing the questions that follow.

How frequently does the level of inter-test variability displayed in the profile occur in the normal population?

0%----------100%

Is inter-test variability associated with neuropsychological dysfunction?

Yes No

If yes:
Taking all test scores into account and assuming that 15 co-normed tests with 32 measures have been administered, what is the approximate dividing point you use in distinguishing between normal and abnormal levels of variability?

Anchor points indicate the maximum discrepancy between the highest and lowest score across a neuropsychological battery in standard deviations.

<0.5 SD	0.5 SD	1.0 SD	1.5 SD	2.0 SD	2.5 SD	3.0 SD	3.5 SD	4.0 SD	4.5 SD	5.0 SD	>6.0 SD

If you would prefer, feel free to provide any comments regarding your responses.

BIBLIOGRAPHY

Ahern, D., Faust, D., & Bridges, A. (in preparation). Impact of sociodemographic adjustments on test score patterns.

Allard G. & Faust, D. (2000). Errors in scoring objective personality tests. *Assessment, 7,* 119-129.

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handal, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84-94.

Binder, L. M., Iverson, G. L, & Brooks, B. L. (2009). To err is human: "Abnormal" neuropsychological scores and variability are common in healthy adults. *Archives of Clinical Neuropsychology, 24,* 31-46.

Bowman, M. L. (2002). The perfidy of percentiles. *Archives of Clinical Neuropsychology, 17*, 295-303.

Brooks, B. L., Strauss, E., Sherman, E. M. S., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology, 50*, 196-209.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G.., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration, scoring, and interpretation* (Rev. ed.). Minneapolis: University of Minnesota Press.

Cohen, J. (1992). A power primer. *Quantitative Methods in Psychology, 112,* 155-159.

Crawford, J. R. & Garthwaite, P. H. (2009). Percentiles please: The case for

expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *The Clinical Neuropsychologist, 23*, 193-204.

Crawford, J. R., Garthwaite, P. H., & Slick, D. J. (2009). On percentile norms in neuropsychology: Proposed reporting standards and methods for quantifying the uncertainty over the percentile ranks of test scores. *The Clinical Neuropsychologist, 23*, 1173-2009.

Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis: University of Minnesota Press.

Faust, D., & Ahern, D. C. (2012). Clinical judgment and prediction. In, D. Faust, *Coping with psychiatric and psychological testimony* (pp.147-208). New York, NY: Oxford University Press.

Hogan, T. P. (2005). 50 widely used psychological tests. In G. P. Koocher, J. C. Norcross, & S. S. Hill III (Eds.), *Psychologists' desk reference* (2nd ed., pp. 101 – 104). New York: Oxford University Press.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York: Oxford University Press.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

National Academy of Neuropsychology, (2003). Connecting in a Virtual World: Listservers in Neuropsychology. *National Academy of Neuropsychology Bulletin, 18*, 6.

The Psychological Corporation. (2002a). *WAIS-III – WMS-III technical manual*. San

Antonio, TX: Author.

The Psychological Corporation. (2002b). *WAIS-III – WMS-III – WIAT-II Writer*. San Antonio, TX: Author.

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20,* 33-65.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (revised). Newbury Park, CA: Sage.

Schofield, W. (1952). Critique of scatter and profile analysis of psychometric data. *Journal of Clinical Psychology, 8,* 16-22.

Schretlen, D. J., & Sullivan, C. (2013). Intraindividual variability in cognitive test performance. In S. Koffler, J. Morgan, I. S. Baron, & M. F. Greiffenstein (Eds.), *Neuropsychology: Science and practice* (Vol. 1, pp. 39-60). New York, NY: Oxford University Press.

Schretlen, D. J., Munro, C. A., Anthony, J. C., & Pearlson, G. D. (2003). Examining the range of normal intraindividual variability in neuropsychological test performance. *Journal of International Neuropsychological Society, 9,* 864-870.

Schretlen, D. J., Testa, S. M., Winicki, J. M., Pearlson, G. D., & Gordon, B. (2008). Frequency and bases of abnormal performance by healthy adults on neuropsychological testing. *Journal of International Neuropsychological Society, 14,* 436-445.

Simons, R., Goddard, R., & Patton, W. (2002). Hand-scoring error rates in

psychological testing. *Assessment, 9,* 292-300.

Spreen, O. & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary* (2rd ed.). New York: Oxford University Press.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford University Press.

Vrieze, S. I. & Grove, W. M. (2009). Survey on the use of clinical and mechanical prediction methods in clinical psychology. *Professional Psychology: Research and Practice, 40,* 525-531.

Wechsler, D. (1997a). *Wechsler Memory Scale – Third Edition: Administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Adult Intelligence Scale – Third Edition: Administration and scoring manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale – Fourth Edition: Administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wedding, D., & Faust, D. (1989). Clinical judgment and decision making in neuropsychology. *Archives of Clinical Neuropsychology*, 4, 233–265.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside Publishing.

Williams, J. E. & Weed, N. C. (2004a). Review of computer-based test interpretation

software for the MMPI-2. *Journal of Personality Assessment*, *83*, 78-83.

Williams, J. E. & Weed, N. C. (2004b). Relative user ratings of MMPI-2 computer-based test interpretations. *Assessment*, *11*, 316-329.